UNITED STATES PATENT APPLICATION FOR:

ADVANCES IN SPIKE ANNEAL PROCESSES FOR ULTRA SHALLOW JUNCTIONS

INVENTORS:

BALASUBRAMANIAN RAMACHANDRAN

RAVI JALLEPALLY

RYAN BOAS

SUNDAR RAMAMURTHY

AMIR AL-BAYATI

HOUDA GRAOUI

JOE SPEAR

ATTORNEY DOCKET NUMBER: AMAT/7916/FEP/OXD/JW

CERTIFICATION OF MAILING UNDER 37 C.F.R. 1.10

I hereby certify that this New Application and the documents referred to as enclosed therein are being deposited with the United States Postal Service on September 22, 2003, in an envelope marked as "Express Mail United States Postal Service", Mailing Label No. EV349851987US, addressed to: Commissioner for Patents, Mail Stop PATENT APPLICATION, P.O. Box 1450, Alexandria, VA 22313-1450.

Signature

Keith M. Tackett
Name

9/22/03
Date of signature

ADVANCES IN SPIKE ANNEAL PROCESSES FOR ULTRA SHALLOW JUNCTIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims benefit of United States provisional patent application serial number 60/412,449, filed September 20, 2002, which is herein incorporated by reference. This application is also related to United States Patent Applications Nos. 10/251,440 [6113], filed September 20, 2002, and 10/267,053 [6519], filed October 7, 2002 [6519]. Each of the aforementioned related patent applications is incorporated by reference herein.

BACKGROUND OF THE INVENTION

Field of the Invention

[0002] This application relates to semiconductor processing technologies, and particularly to a method of annealing semiconductor substrates with rapid thermal processing.

Description of the Related Art

[0003] In today's high speed semiconductor devices, ultra-shallow junctions, low sheet resistance and abrupt lateral junctions are vital to reduce short channel effects and to increase transistor saturation current in source drain extensions. Several techniques have been developed to deal with the issues associated with the formation of shallow, low sheet resistance junctions. Examples of these issues are transient enhanced diffusion (TED), solid solubility, and channeling, which can be resolved by using low energy implants and sharp spike anneals. During low energy implant processes, the implant energies are limited to about 1 keV or less. Thus, TED is minimized because defects caused by the implant processes are confined close to the surface. Sharp spike anneals following the implant processes provide

high dopant activation and effective implant damage removal while minimizing dopant diffusion.

[0004] Spike anneal is typically performed by subjecting a semiconductor substrate having implanted dopants to temperature treatment in a rapid thermal processing (RTP) system. A typical annealing profile using RTP involves ramping up to a target temperature, e.g. 1050 °C, soaking the substrate at the target temperature for a period of time (soak time), and ramping down to a base temperature, e.g. 200 °C. For spike anneal, high ramp rates, e.g., 75 °C/sec or higher, and short (~ 1 sec) or no soak time are desired to prevent excessive dopant diffusion. Besides the tight temperature control requirement, gas composition in the annealing ambient may also need to be controlled. For example, the presence of oxygen has been found to be necessary in order to decrease the evaporation or out-diffusion of implanted dopants such as boron and arsenic, but too much oxygen in the annealing ambient results in oxygen enhanced diffusion (OED) and limits the creation of shallow junctions, particularly when dopants such as boron are used.

[0005] Continued demand for smaller, more compact, faster, and more powerful chips forces the device geometries to scale down to and beyond the 100nm node. Such aggressive downscaling in device geometries increase the Short Channel Effects (SCE). This reduces the differentiation between I_{on} (I_{dsat}) (on state device current which is dependent on device type) and I_{off} (off state device current or leakage currents), which reduction is essential for maintaining the device functionality. Thus the critical challenge in scaling device geometries is to maintain a distinction between I_{on} (I_{dsat}) and I_{off}.

[0006] A key to the challenge in scaling device geometries is in process/performance improvements in Ultra-Shallow Junction (USJ) technology. In a device, I_{on} (I_{dsat}) depends on the amount of active dopant material within the device. Sheet resistance (R_s – as measured by a standard four-point probe method) is one way to measure activation. Higher activation typically provides lower sheet resistance. On the other hand, I_{off} is dependent on the amount of dopant material

that is diffused through the junction. Junction depth is measured as the depth in Angstroms (Å) at which the concentration of the measured species reaches a concentration of 10^{18} atoms/cm^3, as measured by HRD (Dynamic) SIMS (Secondary Ion Mass Spectroscopy) profiles. As junction depth increases, I_{off} increases. Thus, maintaining the differentiation between I_{on} and I_{off} for USJ technology requires a smaller leakage (reduced junction depth) for the same or increased activation (reduced sheet resistance). The sheet resistance and junction depth (X_j) requirements for varying technology nodes are outlined in the International Technology Roadmap for Semiconductors (ITRS), 1999 & 2001 Edition, SIA, San Jose.

[0007] Current USJ technology involves ion implantation followed by a rapid thermal spike annealing process. The main parameters in any spike annealing process are the peak temperature (T_P), and residence time (t_R). A measure of spike sharpness, t_R is defined as the time spent by the substrate within 50 °C of T_P. Higher T_P has the primary effect of causing increased activation, hence causing reduced R_s and increased I_{on}. Different devices have different requirements of activation and hence different choices for T_P. For the same T_P, an increase in residence time has the primary effect of increasing diffusion, hence increasing the leakage currents. Thus, the main effort behind spike anneal is to reduce t_R without compromising on the required level of activation.

[0008] Initial experiments on ramp up rates concluded that increasing ramp up rates greater than 180°C/second did not further improve the sheet resistance and junction depth profiles. Thus, there remains a need for reducing dopant diffusion during annealing of ultra shallow junctions while maintaining high dopant activation.

SUMMARY OF THE INVENTION

[0009] Significant and surprising improvement in reducing dopant diffusion in ultra shallow junctions was obtained, while maintaining high dopant activation, by providing a flow of a carrier gas into the processing chamber and maintaining gas pressure in the processing chamber below 900 Torr, heating the substrate to a peak

temperature above 1000°C, and cooling the substrate at a rate sufficient to provide a residence time less than 1.6 seconds, wherein the residence time is defined as exposure of the substrate to a temperature within 50°C of the peak temperature. In one embodiment, a residence time of 0.9 seconds was obtained for spike annealing of boron doped ultra shallow junctions in a silicon substrate resulting in greater than 10% reduction in boron diffusion.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] Additional objects and features of the invention will be more readily apparent from the following detailed description and appended claims when taken in conjunction with the drawings, in which:

[0011] Figure 1 is a diagrammatic view in vertical cross-section of a portion of an RTP system according to one embodiment of the present invention;

[0012] Figure 2 is a block diagram of a fluid control system that dynamically monitors and controls gas composition and gas pressure in a processing chamber of the RTP system;

[0013] Figure 3 is a flow diagram of a method for annealing a silicon substrate in the processing chamber of the RTP system; and

[0014] Figure 4 is a chart of a heating schedule according to one embodiment of the present invention shown in comparison to a heating schedule shown in a related application.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0015] The present invention demonstrates that spike annealing has been enhanced to meet the ultra-shallow junction requirements for the 100nm-technology node and beyond. An improved spike profile results from hardware and process improvements to the standard Centura Radiance™ RTP chamber as discussed below. The residence time (a metric of spike sharpness) improved to less than 1.6

seconds. In one embodiment, a residence time of 0.9 seconds was achieved resulting in at least a 10% improvement in intrinsic diffusion for boron doped substrates. Intrinsic diffusion modeling suggests that this improvement will translate to nearly 35 % improvement in diffusion of some dopants in ultra shallow junctions.

[0016] Different boron ion implant splits were used to characterize the change in the spike sharpness. Implant splits were optimized to reduce transient enhanced diffusion effects. These substrates were annealed at varying peak temperatures with both improved and non-improved CENTURA® Radiance™ chambers available from Applied Materials, Inc. A greater than 10 % improvement in R_s/X_j performance is observed with the improved chamber for the boron implanted substrates. Such an improvement could translate into greater than 7.5 % improvement in I_{dsat} in various semiconductor devices. These improvements in spike annealing did not have any detrimental effects on Uniformity. Tests comparing R_s averages and Uniformity of the improved chamber with the standard Centura® Radiance spike showed no observable differences.

[0017] The spike annealing of the present invention can be performed in an RTP system capable of maintaining gas pressure in the annealing ambient at a level significantly lower than the atmospheric pressure. An example of such an RTP system is the RADIANCE CENTURA® system commercially available from Applied Materials, Inc., in Santa Clara, California. Figure 1 illustrates an improved rapid thermal processing (RTP) system 10 including a processing chamber 14 for annealing a disk-shaped semiconductor substrate 12, according to one embodiment of the present invention. Chamber 14 is radiatively heated through a water-cooled quartz window 18 by a heating lamp assembly 16. The peripheral edge of substrate 12 is supported by a rotatable support structure 20, which can rotate at a rate of up to about 120 rpm (revolutions per minute). Beneath substrate 12 is a nickel-plated aluminum reflector plate assembly 22 that has an optically reflective coating facing the backside of substrate 12 to enhance the effective emissivity of substrate 12. The optically reflective coating is further described in related application Serial No. 10/267,053, which description is incorporated by reference herein. Reflector plate

assembly 22 is mounted on a water-cooled base 23. Cool down of substrates has been enhanced by increasing the cooling capacity of the water cooled base 23 and by locating the reflector plate assembly 22 closer to the water cooled base 23. Furthermore, the optical coating was enhanced to absorb radiated energy when the lamp assembly is off. Between the top surface of reflector plate assembly 22 and the backside of substrate 12 is a reflective cavity 15.

[0018] In a system designed for processing eight inch (200 mm) silicon wafers, reflector 22 has a diameter of about 8.9 inches, the separation between substrate 12 and the top surface of reflector 22 is about 5-10 mm, and the separation between substrate 12 and the bottom surface of quartz window assembly 18 is about 25 mm. In a system designed for processing twelve-inch (300 mm) silicon wafers, reflector 22 has a diameter of about 13 inches, the separation between substrate 12 and the top surface of reflector 22 is about 18 mm, and the separation between substrate 12 and the bottom surface of quartz window assembly 18 is about 30 mm.

[0019] The temperatures at localized regions of substrate 12 are measured by a plurality of temperature probes 24 that are positioned to measure substrate temperature at different radial locations across the substrate. Temperature probes 24 receive light from inside the processing chamber through optical ports 25, 26, 27, which extend through the top surface of reflector plate assembly 22. While processing system 10 typically may have a total of ten such temperature probes, only some of the probes are shown in Fig. 1. At the reflector plate surface, each optical port may have a diameter of about 0.08 inch. Sapphire light pipes deliver the light received by the optical ports to respective optical detectors (for example, pyrometers), which are used to determine the temperature at the localized regions of substrate 12. Temperature measurements from the optical detectors are received by a first controller 28 that controls the radiative output of heating lamp assembly 16. The resulting feedback loop improves the ability of the processing system to uniformly heat substrate 12.

[0020] During processing, gases for the annealing ambient are introduced into processing chamber 14 through an ambient gas input 30. The ambient gases flow across the top surface of substrate 12 and may react with a heated substrate. Excess ambient gases, as well as any reaction by-products, are withdrawn from processing chamber 14 through an ambient gas output 32 by a pump system 34.

[0021] Most of the excess ambient gases and reaction products can be pumped out of processing chamber 14, but some volatile contaminants, especially those with relatively high vapor pressures such as BO_x and PO_x, may leak into reflective cavity 15 and deposit onto the optical components situated around the reflective cavity. The rate at which volatile contaminants are deposited onto these optical components can be substantially reduced by a flow of a purge gas across the top surface of reflective plate assembly 22. As described in commonly assigned US Patent 6,281,790 B1, which is incorporated herein by reference, a purge fluid injector 40 can be used to produce a substantially laminar flow of a purge gas across the top surface of reflector plate assembly 22.

[0022] The composition of the ambient gases, the flow rate of the purge gas, and the gas pressure in processing chamber 14 are controlled by a fluid control system shown in Fig. 2. In one embodiment of the present invention, the ambient gases comprise oxygen (O_2) and a carrier gas, such as nitrogen (N_2). Mass flow controllers (MFC) 81 and 80 are used to regulate the flow of the carrier gas and oxygen, respectively, into processing chamber 14. A second feedback loop associated with processing chamber 14 controls the oxygen concentration in processing chamber. The second feedback loop includes the MFC 80, an oxygen sensor 95 coupled to the processing chamber 14 and configured to monitor the oxygen concentration in processing chamber, and a second controller 99 coupled between the oxygen sensor 95 and MFC 80, and configured to adjust the MFC based on an oxygen concentration set point (O_2 set point) and the oxygen concentration value detected by the oxygen sensor 95. The second feedback loop insures that a desired O_2 concentration is maintained in processing chamber 14, and may be used as part of a shut down mechanism associated with chamber 14 to prevent substrates from being

processed in chamber 14 when the oxygen concentration cannot be regulated properly.

[0023] When a purge gas, such as nitrogen, is used to prevent deposition of volatile contaminants in the reflective cavity 15, the purge gas is introduced into processing chamber 14 through input 46 which is connected to a filter 86. An MFC 88 is used to regulate the flow of purge gas into processing chamber 14. An adjustable flow restrictor 90 and a mass flow meter (MFM) 92 are used to regulate the rate at which purge gas is removed from processing chamber 14. To reduce the migration of purge gas into the processing region of the processing chamber 14, which is above substrate 12, flow restrictor 90 is adjusted such that the rate at which purge gas is introduced into processing chamber 14 is substantially the same as the rate at which purge gas is removed from processing chamber 14. Solenoid shut-off valves 94 and 96 provide additional control over the flow of purge gas through processing chamber 14.

[0024] A third feedback loop associated with chamber 14 is a closed-loop pressure control system used to regulate the gas pressure in processing chamber 14 by controlling the rate at which gases are removed from processing chamber 14. Still referring to FIG. 2, in one embodiment of the present invention, the pressure control system comprises a pressure control valve 84 at ambient gas output 32, a pressure gauge 98 coupled to processing chamber 14, a programmable logic controller (PLC) 82 coupled to pressure gauge 98, and a third controller 97 coupled between PLC 82 and pressure control valve 84. During the operation of the processing chamber 14, the pressure gauge 98 measures the gas pressure in processing chamber 14 periodically and sends the measured pressure value to PLC 82. The PLC 82 subtracts the measured pressure value from a pressure set point, which indicates the intended gas pressure in chamber 14, and uses an algorithm, such as a proportional integral derivative (PID) control algorithm, to produce a control signal based on a set of tuning parameters. The control signal is then used by PLC 82 to adjust the amount of flow through pressure control valve 84.

[0025] In one embodiment of the present invention, processing chamber 14 is coupled to one or more transfer chambers (not shown), each through a load lock (not shown). The transfer chamber(s) and the associated load lock system facilitate transfers of substrates in and out of processing chamber 14 without substantially changing the gas pressure in processing chamber 14.

[0026] A semiconductor substrate 12, after going through a dopant implant process, can be annealed in processing chamber 14 using a process 300, as illustrated in FIG. 3, according to one embodiment of the present invention. Referring to FIG. 3, before the substrate is loaded into the chamber, processing chamber 14 is pumped down at step 301 to a pressure level between 1 Torr and 900 Torr, preferably a pressure between about 5 Torr and about 300 Torr. Then, while the gas pressure in processing chamber 14 is maintained at step 320 at the pressure level, processing chamber 14 is purged at step 310 with a carrier gas, such as nitrogen, which is introduced into chamber 14 through MFC 81. Other suitable carrier gases include argon, krypton, and xenon. In one embodiment of the present invention, the gas pressure in processing chamber 14 is maintained at step 320 at a level that is in the range of about 5-100 Torr. The flow rate of the carrier gas during the purge step 310 is in the range of about 5-10 standard liter per minute (slm). The purging step reduces the oxygen concentration in processing chamber 14 to below a predetermined minimum value, such as 5 or 50 parts per million (ppm). The time the purging step 310 takes depends on the pressure in processing chamber 14. In one embodiment of the present invention, when the gas pressure in processing chamber 14 is maintained at 10 Torr, it takes less than a few seconds of purging for the oxygen concentration in processing chamber 14 to drop below 5 ppm. At 100 Torr, the purging step may take about 15 seconds, which is still about 4 times quicker than purging at atmospheric pressure, as in the conventional spike anneal process. Also, purging step 310 may not need to be performed for every substrate, as explained below.

[0027] Before or after the purging step 310, substrate 12 is loaded at step 330 into processing chamber 14 from the transfer chamber, which is maintained at near

vacuum and is also purged of oxygen. If substrate 12 is loaded after the purging step, a stabilization step (not shown) may be needed to allow the chamber pressure to stabilize after the loading step 330. Once the chamber pressure is stabilized, while the carrier gas flow is maintained at a predetermined flow rate, such as 5 or 10 standard liters per minute (slm), the substrate is subjected to an improved thermal process at step 350, as described for Fig. 4 below.

[0028] During or shortly before the thermal process step 350, with the flow of the carrier gas continuing, oxygen is introduced at step 360 into processing chamber 14 at a pre-calibrated flow rate through MFC 80. The pre-calibrated oxygen flow rate may depend on the gas pressure in processing chamber 14, the flow rate of the carrier gas, and a predetermined oxygen concentration for the anneal ambient, as discussed above. The desired oxygen concentration for the anneal ambient depends on the type of dopants used, and the performance requirements of the devices being fabricated. A typical oxygen concentration in processing chamber 14 is in the range of 1500 to about 75,000 ppm, and more typically in the range of 10,000 to about 25,000 ppm. After oxygen is introduced into processing chamber 14, the flow rate of oxygen (or the MFC 80) is periodically adjusted by controller 99 based on readings from oxygen sensor 95 so that the predetermined oxygen concentration value is maintained in processing chamber when oxygen in processing chamber is desired. By maintaining the gas pressure in processing chamber 14 at or below 100 Torr, the time it takes for the second feedback loop to adjust the oxygen concentration to the desired value, after a sufficient drift from that value is detected, should be less than a second. This allows accurate and dynamic control of the ambient gas composition during thermal process step 350. The oxygen flow may be turned off at step 370 before the substrate is unloaded at step 380 from processing chamber 14, so as to prevent the oxygen from leaking into the transfer chamber(s).

[0029] Also, with the fast response provided by the low chamber gas pressure, oxygen in chamber 14 can be introduced into processing chamber 14 during a processing phase when a certain level of oxygen concentration in the annealing ambient is desired and can be turned off or down during a processing phase when

oxygen is not desired. In one embodiment of the present invention, oxygen is introduced into processing chamber 14 throughout thermal process step 350. In an alternative embodiment of the present invention, oxygen is introduced into processing chamber 14 only during certain phases of the thermal process step 350. For example, oxygen may be introduced at step 360 into processing chamber near the time when the fast-ramp phase 430 starts and during the soak time (if there is any) in thermal process step 350. Near the time when the substrate starts to cool down, the oxygen flow may be terminated at step 370 either by turning off the MFC 80 or by changing the O_2 set point to zero, allowing the oxygen concentration in the chamber to drop. At sufficiently low pressure, such as 5-20 Torr, the oxygen concentration may drop below the predetermined minimum value before the end of the thermal process step 350.

[0030] A spike annealing process (for T_P = 1050°C) on the non-improved CENTURA® Radiance™ chamber yields a standard residence time of t_S =1.6 seconds, as shown in Fig. 4. Also shown in Fig. 4 is the improved spike profile that resulted from the hardware and process improvements on the CENTURA® Radiance™ chamber discussed above. The improved spike profile increased the spike sharpness by ~40% to provide an improved residence time of t_I = 0.9 seconds. Intrinsic diffusion modeling was also used to quantify the improvements in spike sharpness. An improvement of > 35 % was observed with the "improved spike" profile.

[0031] Referring to Fig. 4, the improved heating schedule of the thermal process step 350 (Fig. 3) shows changes in substrate temperature after a stabilization phase 420, a fast-ramp phase 430, and a cool-down phase 440. During the slow-ramp phase, the substrate is heated slowly and uniformly using open-loop heating to an initial temperature of about 500-600°C. Then the substrate is stabilized at the initial temperature during the stabilization phase 420. Subsequently, during the fast ramp-up phase 430, the substrate is heated again using closed-loop heating so that the substrate temperature rises at a rate of about 50-400°C per second, preferably at least 180°C per second, to a peak temperature of about 1000-1100 °C. The

substrate may be soaked at the peak temperature for a short period of time (<1 second), and is then cooled down during the cool-down phase 440 so that the substrate temperature drops at a rate of about 50-400°C per second, preferably 90°C per second. At the end of thermal process 350, substrate 12 is unloaded at step 380 from processing chamber 14 and another substrate is loaded at step 330 into processing chamber 14 (if more substrates are to be processed).

[0032] The exact order of some of the steps in the process 300 and/or the operation of the processing chamber 14 as described above can be altered. In addition, steps may be added or omitted and process parameters varied depending upon the requirements of a particular processing application and the particular RTP system in which the annealing process takes place. The above operations and the order in which they are presented are chosen for illustrative purposes and to provide a picture of a complete run sequence.

Examples

[0033] All comparison experiments were performed on B-ion implanted substrates that were pre-amorphised with Ge to reduce deep implants depths and enhance comparison of implant depths.

[0034] B-ion implants were performed using an Applied Materials Quantum™ LEAP implanter. Silicon substrates were pre-amorphised (PAI) with Ge, prior to the B implantation. The Ge pre-amorphisation was optimized for each B implant condition. The B implant energy ranged from 200 to 500 eV, and the implanted dose ranged from 10^{15} to $2x10^{15}$ atoms/cm^2. The initial set of experiments also included 1 keV boron without PAI (dose = 10^{15} atoms/cm^2). All boron implants were performed in "decelerated" mode: B ions were extracted at 2 keV and decelerated to the desired final energy level. These substrates were then annealed with varying peak temperatures, using the improved spike profile of Fig. 4. The annealed substrates were measured for sheet resistance and analyzed for junction depth by characterizing boron profiles in silicon through dynamic Secondary Ion Mass Spectroscopy (SIMS) analysis.

[0035] A specific implant condition (Boron, 500eV/10^{15} atoms/cm^2, with Ge PAI) was also studied for its susceptibility to diffuse at lower temperatures. A spike anneal was performed on these substrates at T_P <1000°C. Subsequently these substrates were analyzed for sheet resistance and junction depth.

[0036] Low spike peak temperature (T_P) experiments with 500 eV, 10^{15} atoms/cm^2 B-implanted substrates with Ge-PAI showed little or no significant diffusion. The observed activation was also very low (high R_s). Since there is no appreciable diffusion/activation at low temperatures, it can be concluded that for these particular implant conditions, diffusion occurs only at high temperatures. For the above Boron implanted substrates, it is unlikely to perceive any change in improving spike profiles at T_P < 1000°C.

[0037] Remaining substrates were annealed with the standard t_S-spike of 1.6 seconds and the improved t_I-spike of 0.9 seconds, using a wide range of spike peak temperatures from 1000°C to 1100°C. An improvement of greater than 10 % was attained by using the improved t_I-spike profile over the standard t_S-spike. Also, the improved spike profile compares favorably with available data on spike anneal processes using other RTP technologies.

[0038] Experiments were conducted with the improved spike profile to understand the impact on the temperature uniformity across the substrate. This was done by comparing the uniformity using the improved spike profile with the uniformity using the standard spike profile. No difference in Uniformity between the standard spike and improved spike was observed. Thus, the given improvements in spike annealing did not impact temperature uniformity across the substrate.

[0039] While the foregoing is directed to embodiments of the present invention, other and further embodiments of the invention may be devised without departing from the basic scope thereof, and the scope thereof is determined by the claims that follow.